Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Member Name: • Shaik Ahmad Basha ahmadshaik982basha@gmail.com Contribution: • Exploring Data • Data Wrangling • Data Cleaning • Checking for Null Values and Duplicated Values • Analyze How Numerical Features Varies with Dependent Feature • Analyze How Categorical Features Varies with Dependent Feature • Performed EDA on Dataset • Identifying and Removing Outliers • Removing Multicollinearity • Feature Scaling • Fitting the Data into Various Regression Models like Linear Regression, Lasso Regression, Ridge Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, XGBoost Regressor. • Plotting Feature Importance. • Tuning the Hyperparameters to Avoid Overfitting
Please paste the GitHub Repo link.
GitHub Link: - https://github.com/ahmedshaik982/Bike_Sharing_Demand_Prediction
Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Problem Statement:

Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

Our main objective behind this project is to explore and analyze the data to discover the key understandings. And to predict the count of bikes required at each hour by using regression models

Approaches:

The first step imported all the necessary libraries like NumPy, Pandas etc. and then collected the data. I started with understanding the data like what are the columns and their meanings and data types.

After that, the second step is data preprocessing. Data preprocessing is a process where raw data is converted into clean data. The dataset has no null values and duplicated values. The Third step is to analyze the data using Exploratory data analysis techniques.

And then I moved on to Feature engineering and preprocessing where I removed outliers present in the data and I also removed multicollinearity. I also scaled the input features by using MinMaxScaler.

And the last step is fitting the data into various regression models and evaluating the metrics of test dataset and plotting feature importance.

Conclusions:

Most number of bikes are rented in the 18th Hour followed by 19th hour. And the least is 4th hour. In summer season, most number of bikes are rented and the least is winter season. In working days(No Holiday), most number of bikes are rented. Thursday has high count of rented bike. In a functioning day, most number of bikes are rented.

The distribution between numerical features and dependent feature (Rented Bike Count) has been spread out entire area which means there is no specific relation between them.

After fitting the data into various regression models, we can conclude that
Tree based models performs well than linear models because, the independent features are not linearly related to the dependent feature ('Rented Bike Count')

The optimal model is XGboost regressor and Gradient boosting regressor because those only have required score(no overfitting and no underfitting).